Bot Meets Shortcut: How Can LLMs Aid in Handling Unknown Invariance OOD Scenarios?

Shiyan Zheng[1, 2], **Herun Wan**[1], **Minnan Luo**[1, 2*], **Junhang Huang**[3]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]State Key Laboratory of Communication Content Cognition, China
[3]Beijing Institute of Technology, China
2223515385@stu.xjtu.edu.cn, minnluo@xjtu.edu.cn

Abstract

While existing social bot detectors perform well on benchmarks, their robustness across diverse real-world scenarios remains limited due to unclear ground truth and varied misleading cues. In particular, the impact of shortcut learning, where models rely on spurious correlations instead of capturing causal task-relevant features, has received limited attention. To address this gap, we conduct an in-depth study to assess how detectors are influenced by potential shortcuts based on textual features, which are most susceptible to manipulation by social bots. We design a series of shortcut scenarios by constructing spurious associations between user labels and superficial textual cues to evaluate model robustness. Results show that shifts in irrelevant feature distributions significantly degrade social bot detector performance, with an average relative accuracy drop of 32% in the baseline models. To tackle this challenge, we propose mitigation strategies based on large language models, leveraging counterfactual data augmentation. These methods mitigate the problem from data and model perspectives across three levels, including data distribution at both the individual user text and overall dataset levels, as well as the model's ability to extract causal information. Our strategies achieve an average relative performance improvement of 56% under shortcut scenarios.

Code — https://github.com/worfsmile/BotsMeetShortcut

Extended version — https://arxiv.org/abs/2511.08455

1 Introduction

Social bot detection has become a significant research topic because of the rapid development of social networks. By consensus, social bots are automated accounts controlled by computer programs that mimic human behavior on social platforms (Ferrara et al. 2016; Cresci 2020). These accounts perform actions such as posting content, commenting, liking, and sharing, actively participating in digital social interactions. Given their substantial influence on information dissemination and public opinion shaping, social bots have attracted increasing attention from both academia and society (Elmas, Overdorf, and Aberer 2022). With the continued advancement of social bot detection research, many deep learning based methods have achieved increasingly strong performance on benchmark datasets (Feng et al. 2022; Liu et al. 2023; Yang et al. 2024; Li et al. 2025).

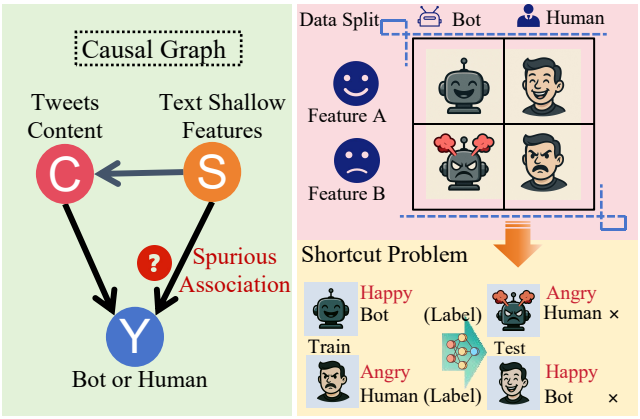

Figure 1: Schematic illustration of the shortcut scenario. As shown on the left, the causal graph depicts how spurious features (*i.e.*, shortcuts) may interfere with inference, leading the model to learn incorrect reasoning from the training set. For instance, on the right, the users are partitioned after associating task-irrelevant feature (*e.g.*, sentiment) with their labels. As a result, the detector fails to generalize, and tends to make incorrect predictions when evaluated on diverse test instances.

However, social bot detection task has long been considered an ongoing arms race (Cresci 2020; Feng et al. 2024). Social bots are not static adversaries but continuously evolve to evade detection, from early attempts to obscure user-level metadata (Yang et al. 2020), to mimicking human-like language and retweeting genuine content (Feng et al. 2024), and more recently, to simulating complex social behaviors such as realistic follow networks and strategic user interactions (Feng et al. 2021b). This evolutionary trajectory demonstrates the bots' strong adaptability and adversarial nature, posing continuous challenges to static detection methods and underscoring the need for more robust and generalizable solutions (Cresci et al. 2023).

What's more, despite the impressive performance of state-of-the-art models, numerous studies suggest that deep learn-

*Corresponding author

ing models often exploit spurious correlations or shortcuts, which are shallow and irrelevant features correlated with task labels but lacking causal relevance (fig. 1). This hinders their ability to learn truly meaningful representations (Geirhos et al. 2020; Wan et al. 2025), raising concerns about their robustness and generalization. This challenge is particularly pronounced in the social bot detection task, as current detectors still struggle to generalize across different benchmarks, data distributions, and time due to biases in the datasets (Hays et al. 2023). Many models are closely tied to the specific datasets or network structures used during training (Li et al. 2025), limiting their effectiveness in dynamic, real-world scenarios. This challenge is largely driven by dataset biases and the ever-evolving nature of social network structures, discourse topics, and user behavior (Cresci 2020; Cresci et al. 2023), prompting a growing body of research to focus on this issue (Mannocci et al. 2024; Tardelli et al. 2024).

In this work, we conduct an in-depth investigation into the generalization ability of social bot detection models to evaluate how detectors are influenced by inherent shortcuts. Focusing on the intrinsic textual features of social media users, we design a series of distribution shift scenarios to examine whether shallow **text-level** perturbations can trigger shortcut learning in existing detection systems (Geirhos et al. 2020; Wan et al. 2025). Building on these insights, we further propose a set of debiasing strategies both at the data and model levels, leveraging large language models (LLMs) to enhance robustness under unknown invariance conditions. Our main contributions are twofold:

- **Potential Shortcut in Social Bot Detection.** We first investigate the shortcut learning problem in social bot detection by focusing on endogenous textual features, namely sentiment, topic, emotion, and human values. Inspired by the concept of spurious correlations in distribution shifts and shortcut learning (Lu et al. 2019; Geirhos et al. 2020), we align labels with these superficial attributes to construct pseudo-correlated, biased scenarios. We partition the data based on textual feature to create train set and test set across three most authoritative and widely-used social bot detection datasets, Cresci-2015-Data (Cresci et al. 2015), Cresci-2017-Data (Cresci et al. 2017), and Twibot-20 (Feng et al. 2021a). As the most commonly used detecting methods can be broadly categorized into text-based (Feng et al. 2024) and graph-based (Feng et al. 2021b) approaches, we evaluate the performance of representative baseline models under the given conditions. Our experimental results show a relative performance drop averaging **33%** for text-based models and **30%** for graph-based models, while commonly used debiasing methods remain largely ineffective in mitigating the degradation caused by such biases.
- **Mitigation Methods.** To address this challenge, we further explore counterfactual intervention approaches using LLMs (Liu, Kusner, and Blunsom 2021; Mishra et al. 2024) to generate **rewritten** text. Our proposed methods target bias reduction from three perspectives: the surface tendency of user text features, the construction of augmented datasets, and the representation ability of the language feature extraction model. Through these insights, we propose targeted and effective mitigation strategies to enhance robustness in the presence of spurious correlations, achieving an average relative performance improvement of **59%** on text-based models and **53%** on graph-based models compared with the shortcut setting before augmentation in our observations.

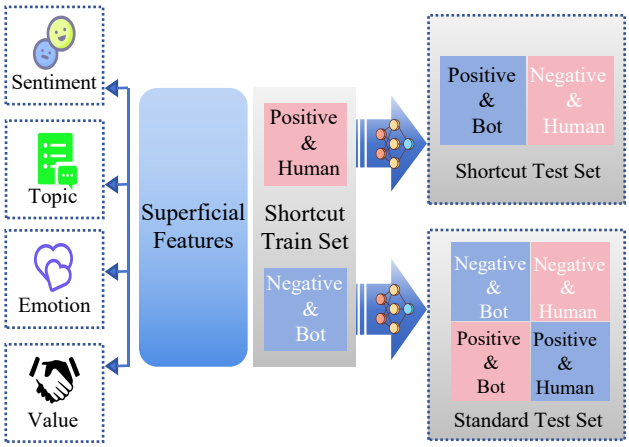

Figure 2: A diagram of our shortcut settings. We focus on the superficial features of the text, such as sentiment, topic, emotion, and human values, and set shortcuts to these features in the training set. In the test set, we either reverse the pseudo-correlation between features and the label in the shortcut test set or eliminate these shortcuts in the standard test set.

2 Potential Shortcuts

Shortcut learning occurs when models exploit superficial features (*e.g.*, sentiment cues in the social bot detection), which are often shallow and easier to capture (Geirhos et al. 2020). In the social bot detection task, the classification criteria are quite complex and therefore easily influenced by unrelated factors. We explore potential shortcut learning factors in **user text**, which is the most information-rich and bot-manipulable medium in social networks, and empirically demonstrate that social bot detection can be largely affected by shortcut learning through constructing endogenous shortcut scenarios (fig. 2).

2.1 Shortcut Learning Scenarios Setup

Formulate the social bot detection as a binary classification task in which each instance corresponds to a user $u \in \mathcal{U}$ where $\mathcal{U}$ denotes the set of all users connected through the social network graph structure. Users are represented by the textual feature vector extracted from their selected posts and each instance has an associated label $y \in \{0, 1\}$, indicating whether the user is a human $(y = 0)$ or a bot $(y = 1)$. To analyze user's textual features from different perspectives, we denote $\phi^{\text{causal}}_{\text{task}}(u)$ as the semantic signals that are causally related to the user label y, while $\phi^{\text{spu}}_{\text{fea}}(u)$ indicates the textual

RoBERTa		Cresci-2015-Data		Cresci-2017-Data		Twibot-20	
		$\text{Shortcut}_{\text{te}}$	$\text{Standard}_{\text{te}}$	$\text{Shortcut}_{\text{te}}$	$\text{Standard}_{\text{te}}$	$\text{Shortcut}_{\text{te}}$	$\text{Standard}_{\text{te}}$
Sentiments	$\text{Standard}_{\text{tr}}$	.945	.955	.884	.900	.682	.685
	$\text{Shortcut}_{\text{tr}}$	.565	.780	.287	.625	.051	.523
	Difference	40%↓	18%↓	67%↓	30%↓	92%↓	23%↓
Emotions	$\text{Standard}_{\text{tr}}$	.982	.979	.963	.940	.684	.687
	$\text{Shortcut}_{\text{tr}}$	.849	.917	.555	.766	.110	.520
	Difference	13%↓	6%↓	42%↓	18%↓	83%↓	24%↓
Topics	$\text{Standard}_{\text{tr}}$	.985	.973	.894	.885	.663	.684
	$\text{Shortcut}_{\text{tr}}$	.941	.962	.587	.757	.164	.535
	Difference	4%↓	1%↓	34%↓	14%↓	75%↓	21%↓
Values	$\text{Standard}_{\text{tr}}$	.915	.928	.890	.894	.691	.679
	$\text{Shortcut}_{\text{tr}}$	.659	.827	.514	.755	.180	.550
	Difference	28%↓	10%↓	42%↓	15%↓	73%↓	18%↓

Table 1: Accuracy of RoBERTa under standard and shortcut training and testing conditions. The vertical dimension denotes training configurations, $\text{Standard}_{\text{tr}}$ (randomly sampled, balanced) versus $\text{Shortcut}_{\text{tr}}$ (spuriously correlated shallow features), while the horizontal dimension denotes test feature distributions, $\text{Standard}_{\text{te}}$ (randomly sampled, balanced) versus $\text{Shortcut}_{\text{te}}$ (distributional shift). The difference between paired entries reported as the relative drop from standard to shortcut settings reflects the model's vulnerability to spurious correlations.

attribute associated with certain spurious factors (*e.g.*, text emotion, topic and so on) that may correlate with y while lacking causal relevance.

Let $\mathcal{S}_{\text{fea}}$ denote the set of possible values of $\phi^{\text{spu}}_{\text{fea}}(u)$ associated with some specific feature (*e.g.*, emotion) and partition this set into two mutually exclusive and internally similar subsets: $\mathcal{S}^{\text{pos}}_{\text{fea}}$ and $\mathcal{S}^{\text{neg}}_{\text{fea}}$. Based on this partition, we define two corresponding instance sets as

$$\mathcal{U}^{\text{pos}}_{\text{fea}} = \{u \mid \phi^{\text{spu}}_{\text{fea}}(u) \in \mathcal{S}^{\text{pos}}_{\text{fea}}\},$$
$$\mathcal{U}^{\text{neg}}_{\text{fea}} = \{u \mid \phi^{\text{spu}}_{\text{fea}}(u) \in \mathcal{S}^{\text{neg}}_{\text{fea}}\}.$$

We collect the instances into set $\mathcal{D} = \{(u_i, y_i)\}_{i=1}^{n}$ and split it as the **Shortcut Train Set** $\mathcal{D}^{\text{str}}_{\text{fea}}$ and **Shortcut Test Set** $\mathcal{D}^{\text{ste}}_{\text{fea}}$ based on the presence of special *spurious feature*, *i.e.*,

$$\begin{aligned}\mathcal{D}^{\text{str}}_{\text{fea}} =& \{(u_i, y_i) \in D \mid u_i \in \mathcal{U}^{\text{pos}}_{\text{fea}},\ y_i = 1\} \cup \\ & \{(u_i, y_i) \in D \mid u_i \in \mathcal{U}^{\text{neg}}_{\text{fea}},\ y_i = 0\}, \\ \mathcal{D}^{\text{ste}}_{\text{fea}} =& \{(u_i, y_i) \in D \mid u_i \in \mathcal{U}^{\text{pos}}_{\text{fea}},\ y_i = 0\} \cup \\ & \{(u_i, y_i) \in D \mid u_i \in \mathcal{U}^{\text{neg}}_{\text{fea}},\ y_i = 1\}.\end{aligned}$$

For example, one may train on happy bots (sampled from $\mathcal{U}^{\text{pos}}_{\text{emotion}}$ labeled 1) and angry humans (sampled from $\mathcal{U}^{\text{neg}}_{\text{emotion}}$ labeled 0), then test on angry bots and happy humans in the **Shortcut Test Set**. Note that in the **Standard Test Set**, labels are *independent* of shortcut features.

In this work, we specifically focus on several types of shallow textual features that could potentially trigger shortcut learning: sentiment, topic, emotion, and human values. For instance, in the case of sentiment, we assign users whose texts express a clearly *positive* tone to $\mathcal{U}^{\text{pos}}_{\text{senti}}$, and those with a *negative* tone to $\mathcal{U}^{\text{neg}}_{\text{senti}}$. For topic, those texts discussing *daily life* are assigned to $\mathcal{U}^{\text{pos}}_{\text{topic}}$, while those pertaining to *pop culture* and *sports* fall into $\mathcal{U}^{\text{neg}}_{\text{topic}}$. Specifically, we assign users whose texts corresponding feature types cannot be clearly categorized into the above-defined sets to $\mathcal{U}^{\text{neu}}_{\text{fea}}$. Based on this subset division, we obtain train set and test set accordingly to construct our shortcut learning setting through text filtering (more details are provided in appendix C).

2.2 Impact on Classifiers

We test the impact of the above-mentioned scenarios on the detectors. We use some of the most well-known and widely used social bot detection datasets, Cresci-2015-Data (Cresci et al. 2015), Cresci-2017-Data (Cresci et al. 2017), and Twibot-20 (Feng et al. 2021a). For these datasets, we design shortcut learning scenarios and evaluate three different approaches:

- **Based on Language Models (LM):** We use the RoBERTa (Liu et al. 2019) model with frozen parameters to generate feature embeddings for the user's text. We then use an MLP classifier to classify the text.
- **Based on Graph:** We use the embedded text as features and incorporate social network graph structure information, applying BotRGCN (Feng et al. 2021b) classifier for classification.
- **Based on Debiasing Approaches:** We try to employ debiasing methods to reduce the effect of shortcut features on classification. We tested a representative causal decoupling-based debiasing model, CIGA (Chen et al. 2022), by representing tweets with abstract meaning representations (AMR) (Banarescu et al. 2013).

The experimental results, as shown in table 1 and appendix D, indicate that all of these methods were affected by the shortcut learning scenario. For example, in the text-based method, compared to the normal distribution scenario (*i.e.*, training on the standard train set), the relative classification accuracy in *shortcut test setting* on average decreases by **50%**, with the largest decrease being **92%**. In *standard test*

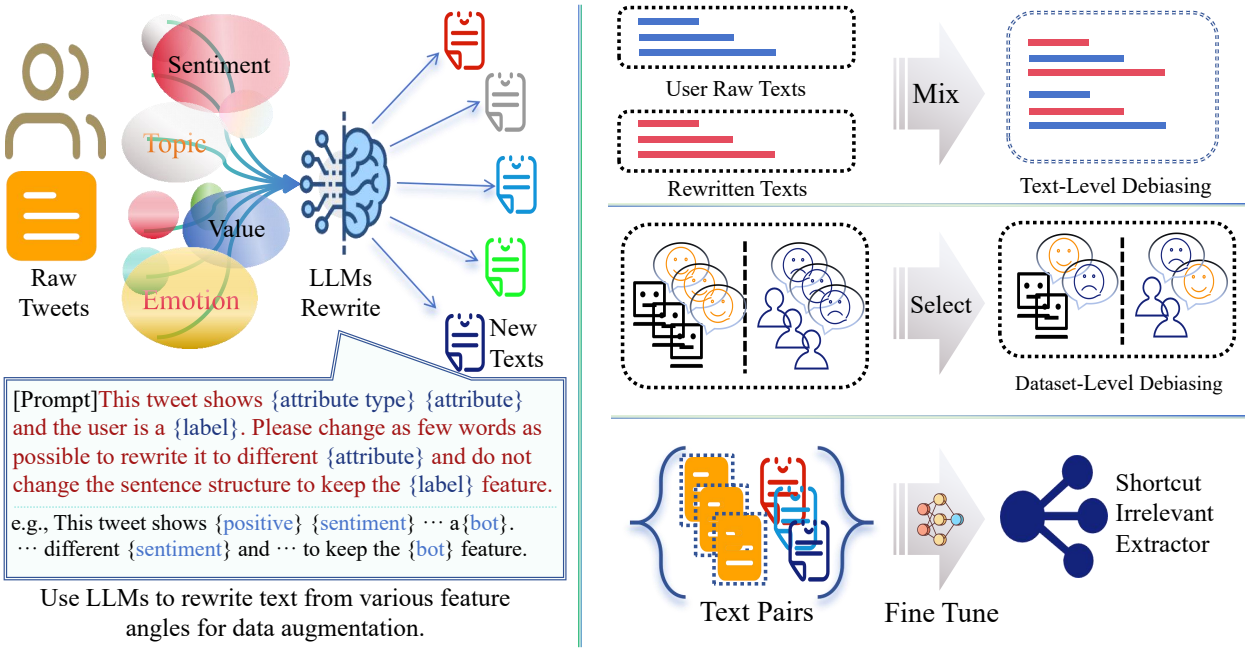

Figure 3: Overview of our shortcut learning mitigation framework. The left side illustrates how LLMs are used to augment data by rewriting text from different attribute perspectives, while preserving the user's label-related semantics. The right part demonstrates the mitigation process at three levels: balancing the semantic content of individual users' texts at the user level, balancing feature distributions across classes at the dataset level, and enhancing the feature extractor's ability to capture causal information across different shortcut shifts at the language model embedding level by employing contrastive learning.

setting, the average relative accuracy drop is **17%**, with the most significant decrease being **30%**, which demonstrates that existing model architectures suffer significant performance degradation. And the debiasing model (we implement it by combining AMR's (Banarescu et al. 2013) explicit semantic graphs with CIGA's (Chen et al. 2022) causal disentanglement) is not effective enough to alleviate this issue, as it exhibits poor performance in the shortcut scenarios (compared in table 2 and detail in appendix D.

3 Mitigation Methods

Leveraging the competency of LLMs in social bot detection tasks (Feng et al. 2024), we employ a counterfactual data augmentation (CDA) strategy comprising two main components: (1) At the first step, we perform CDA on the **Training Set** to generate texts that reverse specific biased feature while preserving the main semantic content and label criteria. (2) We mitigate shortcut-inducing correlations at three levels, including semantic patterns in the *text level*, skewed feature distributions in the *dataset level*, and embedding language model's debiasing ability in the *model level* (fig. 3).

3.1 Counterfactual Data Augmentation

Specialized LLMs can identify potential superficial feature biases **in training data** by analyzing text patterns (*e.g.*, clearly positive tone). Then to counteract these spurious correlations, we employ a prompt-based LLM **rewriting** method (using DeepSeek API (Guo et al. 2025) in our implementation) to generate counterfactual text while maintaining semantic consistency. Specifically, we prompt the model to alter the expression of specified shallow features (*e.g.*, sentiment, human values and topic cues) *without* changing the core meaning and sentence structure of the original texts, while preserving the distinguishing characteristics between social bots and humans in the original texts (some examples in appendix E). And by rewriting an original text T_{raw} into a new version T_{new}, we obtain a text pair $(T_{\text{raw}}, T_{\text{new}})$.

3.2 Text-Level Debiasing

Each user u is associated with a set of tweets $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ ($k = 5$ in our experiments), and suppose these texts exhibit a specific bias under a shallow feature attribute. We prompt the LLM to rewrite each tweet such that the resulting set $\mathcal{T}' = \{T'_1, T'_2, \ldots, T'_k\}$ presents a different or neutral tendency with respect to the attribute, while preserving the original label inference basis. We then combine original and rewritten texts to select a balanced subset with minimized bias. To this end, we define a feature bias score as

$$f = \left| \frac{e^{R_{\text{pos}}} - e^{R_{\text{neg}}}}{e^{R_{\text{neu}}}} \right|, \tag{1}$$

where R_{pos}, R_{neg}, and R_{neu} represent proportions of tweets with positive, negative expressions of the attribute, and R_{neu} is proportion of types out of them (*i.e.,* defined in section 2.1). We evaluate all combinations of original/revised

RoBERTa		Cresci-2015-Data		Cresci-2017-Data		Twibot-20	
		$\text{Shortcut}_{\text{te}}$	$\text{Standard}_{\text{te}}$	$\text{Shortcut}_{\text{te}}$	$\text{Standard}_{\text{te}}$	$\text{Shortcut}_{\text{te}}$	$\text{Standard}_{\text{te}}$
Sentiments	$\text{Shortcut}_{\text{tr}}$	.565	.780	.287	.625	.051	.523
	AMR+CIGA	.678 (19%↑)	.838 (7%↑)	.271 (5%↓)	.671 (7%↑)	.112 (120%↑)	.531 (1%↑)
	Text-Level*	.840 (48%↑)	.910 (16%↑)	.597 (108%↑)	.757 (21%↑)	.229 (349%↑)	.559 (6%↑)
	Dataset-Level*	.835 (47%↑)	.912 (16%↑)	.667 (132%↑)	.794 (27%↑)	.288 (466%↑)	.571 (9%↑)
	Model-Level*	.864 (52%↑)	.907 (16%↑)	.690 (140%↑)	.836 (33%↑)	.415 (715%↑)	.580 (10%↑)
Emotions	$\text{Shortcut}_{\text{tr}}$	.849	.917	.555	.766	.110	.520
	AMR+CIGA	.770 (9%↓)	.870 (5%↓)	.382 (31%↓)	.679 (11%↓)	.232 (109%↑)	.538 (3%↑)
	Text-Level*	.936 (10%↑)	.964 (5%↑)	.670 (20%↑)	.819 (6%↑)	.235 (113%↑)	.569 (9%↑)
	Dataset-Level*	.936 (10%↑)	.971 (5%↑)	.610 (9%↑)	.782 (2%↑)	.288 (160%↑)	.569 (9%↑)
	Model-Level*	.923 (8%↑)	.933 (1%↑)	.628 (13%↑)	.803 (4%↑)	.400 (262%↑)	.585 (12%↑)
Topics	$\text{Shortcut}_{\text{tr}}$	.941	.962	.587	.757	.164	.535
	AMR+CIGA	.774 (17%↓)	.850 (11%↓)	.405 (30%↓)	.656 (13%↓)	.189 (15%↑)	.530 (0%↓)
	Text-Level*	.979 (4%↑)	.972 (1%↑)	.780 (32%↑)	.833 (9%↑)	.296 (80%↑)	.577 (7%↑)
	Dataset-Level*	.992 (5%↑)	.962 (0%↓)	.807 (37%↑)	.807 (6%↑)	.607 (270%↑)	.567 (5%↑)
	Model-Level*	.954 (1%↑)	.953 (0%↓)	.839 (42%↑)	.867 (14%↑)	.522 (218%↑)	.635 (18%↑)
Values	$\text{Shortcut}_{\text{tr}}$	.659	.827	.514	.755	.180	.550
	AMR+CIGA	.592 (10%↓)	.785 (5%↓)	.572 (11%↑)	.739 (2%↓)	.225 (24%↑)	.536 (2%↓)
	Text-Level*	.762 (15%↑)	.873 (5%↑)	.711 (38%↑)	.826 (9%↑)	.295 (63%↑)	.567 (3%↑)
	Dataset-Level*	.633 (3%↓)	.818 (1%↓)	.771 (50%↑)	.828 (9%↑)	.407 (126%↑)	.595 (8%↑)
	Model-Level*	.615 (6%↓)	.812 (1%↓)	.830 (61%↑)	.869 (15%↑)	.488 (171%↑)	.614 (11%↑)

Table 2: Mitigation effects of our methods on RoBERTa model (* indicates our strategies). We compare our mitigation strategies against the original baseline and representative debiasing methods under the shortcut setting, and report the relative improvement over the original shortcut scenario. The best performance in each group is highlighted with an underline. Results demonstrate that our methods effectively alleviate the impact of shortcuts, and in most cases, the performance approaches or even matches that of the standard setting.

texts (using binary selection vector of size k), and select the one minimizing f to obtain a debiased tweet set $\mathcal{T}'' = \{T_1'', T_2'', \ldots, T_k''\}$ for each user (algorithm 1 in appendix F).

3.3 Dataset-Level Debiasing

After obtaining the bias-mitigated texts at the user level, we also conduct dataset-level debiasing to reduce the shortcut learning between labels and superficial features.

For each label class $y \in \{0, 1\}$, we randomly divide its corresponding samples in the training set: half use the original tweets $\mathcal{T}$, and the other half use the rewritten tweets $\mathcal{T}'$. Due to the limitations of LLMs in precisely controlling rewriting directions and the inherent directional bias in the original text, not all samples can be reliably modified as expected, meaning that not all users can obtain $\mathcal{T}'$. In cases where the samples cannot be evenly split, the remaining instance is assigned the mixed version $\mathcal{T}''$. This strategy ensures a relatively balanced distribution of shallow features across different classes and weakens the model's reliance on spurious correlations (algorithm 2 in appendix F).

3.4 Model-Level Mitigation

On top of getting counterfactual texts, we further fine-tune the language feature extractor using contrastive learning to enhance its ability to capture causal information.

Data Preparing From the CDA steps, we obtain a collection of text pairs $\{(T_{\text{raw}}, T_{\text{new}})\}$, where each pair consists of an original text and its rewritten version. And we collected over 200,000 such pairs, which serve as the dataset for our fine-tuning process.

Debiasing via Contrastive Fine-Tuning To remove shortcut features from text embeddings, we fine-tune a pretrained language feature extractor model M (we use RoBERTa (Liu et al. 2019) in our experiment) using a joint objective that encourages manifold preservation and suppresses spurious information. Denote M_{raw} as the model before finetuning, and M_{finetune} as the model after finetuning. Given a text pair $(T_{\text{raw}}, T_{\text{new}})$, we compute the embedding of the raw text as $h_{\text{raw}} = M_{\text{raw}}(T_{\text{raw}})$, $h_{\text{pos}} = M_{\text{finetune}}(T_{\text{raw}})$, and, where applicable, a contrastive example $h_{\text{neg}} = M_{\text{finetune}}(T_{\text{new}})$. The overall loss to be minimized is defined as

$$\mathcal{L} = \mathcal{L}_{\text{manifold}} + \lambda \mathcal{L}_{\text{MI}}. \tag{2}$$

where λ is a hyperparameter that balances the contribution of the mutual information loss relative to the manifold preservation loss.

The manifold loss preserves semantic structure across texts by aligning the similarity distributions between raw and transformed embeddings, that is a modified version of

that introduced by Park et al. (2019), *i.e.*,

$$\mathcal{L}_{\text{manifold}} = \lambda_1 \mathcal{L}_{\text{positive}} + \lambda_2 \mathcal{L}_{\text{negative}}, \quad (3)$$

$$\mathcal{L}_{\text{positive}} = KL(\text{Sim}(H_{\text{pos}}), \text{Sim}(H_{\text{raw}})), \quad (4)$$

$$\mathcal{L}_{\text{negative}} = KL(\text{Sim}(H_{\text{neg}}), \text{Sim}(H_{\text{raw}})), \quad (5)$$

where λ_1, λ_2 denotes a hyperparameter. $KL(\cdot, \cdot)$ denotes the Kullback-Leibler (KL) divergence. And $\text{Sim}(H)$ denotes cosine similarities computed between h representations within a batch.

To remove shortcut signals, we maximize the mutual information (MI) between positive and negative examples, *i.e.*,

$$\mathcal{L}_{\text{MI}} = -I(h_{\text{pos}}, h_{\text{neg}}). \quad (6)$$

We maximize mutual information by maximizing its lower bound, utilizing the InfoNCE method, shown in eq. (7) proposed by Oord, Li, and Vinyals (2018).

$$I_{\text{NCE}} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f(h^i_{\text{pos}}, h^j_{\text{neg}})}}{\frac{1}{N} \sum_{j=1}^{N} e^{f(h^i_{\text{pos}}, h^j_{\text{neg}})}} \quad (7)$$

$$= \frac{1}{N} \sum_{i=1}^{N} f(h^i_{\text{pos}}, h^i_{\text{neg}}) - \frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{1}{N} \sum_{j=1}^{N} e^{f(h^i_{\text{pos}}, h^j_{\text{neg}})} \right].$$

Here, f is a learnable non-negative scoring function, and I_{NCE} serves as a variational lower bound of the mutual information I, with N related to the batch size. This approach maximizes mutual information by estimating a lower bound.

This training process yields a language feature extraction model that preserves semantic structure while reducing spurious correlations in the embedding space. And we select the finetuned model with a batch size of 256 and 1000 training steps for this task. Some discussion on the performance of finetuning can be found in appendix H .

3.5 Mitigation Methods Performance

Our strategy effectively mitigates the issue in both *text-based* (table 2) and *graph-based* settings (appendix G). Experimental results show that in the *text-based* scenario, our method achieves an average relative improvement of **59%** over the original model. In the *graph-based* scenario, the average relative improvement reaches **53%**. These results underscore the effectiveness of our strategy.

4 Discussions

4.1 Gap in the Exploration of Potential Shortcuts

In our analysis, we observed a significant drop in model performance when the distribution of shallow textual features is altered. On the one hand, previous studies have shown that distributional biases in lexical usage, sentiment, and topic preference exist between social bot users and human users in commonly used datasets (Li et al. 2025). And by leveraging these biases, some approaches have achieved impressive performance in the social bot detection task. However, we argue that such performance gains are not inherently robust in our experiment.

On the other hand, we emphasize that models should not depend on these shortcuts, especially given that in real-world scenarios, the distributions of such features are often more balanced between user types. As shown in fig. 4, the topic distribution in benchmark dataset such as Cresci-2017-Data exhibits little divergence between humans and bots. This highlights the risk of shortcut learning and motivates the need for models that generalize beyond surface-level cues.

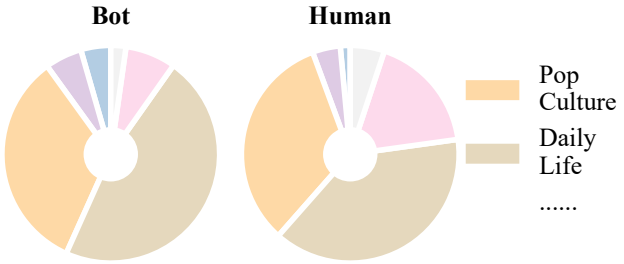

Figure 4: The topic distribution across different tags in the Cresci-2017-Data. There is no significant difference between human and bot distributions, indicating that topic features *should not* be exploited as cues by detectors.

4.2 Shortcut Scenarios Impact on Model

In addition to demonstrating that shortcut scenarios significantly degrade model accuracy, we further examine how such features affect model confidence. Following the approach of Guo et al. (2017), we compute the expected calibration error (ECE) (the lower the better) and average prediction confidence.

We observe that when training stage occurs shortcut features, the model's accuracy drops, yet its prediction confidence increases notably in some cases. This suggests that the model becomes more overconfident despite making more errors, indicating a lack of proper uncertainty calibration in the presence of shortcut learning.

4.3 The Appropriateness of Mitigation Strategies

In prior work, Feng et al. (2024) demonstrated that LLMs can effectively distinguish between social bots and human users in detection tasks, thereby ensuring that our prompt-based rewriting preserves the original text labels. To further validate the effectiveness of LLMs in mitigating shortcut reliance, we evaluate the semantic consistency between the original and rewritten texts, confirming that only the targeted shallow features have been modified. We compute the edit distance at the token level and the cosine similarity at the embedding level between each pair of original and modified texts, and employ KDE to illustrate the overall distribution.(fig. 6). The rewriting yields a token-level edit similarity of over 0.7 (substantially higher than the average similarity of 0.03 observed in over 1 million randomly paired texts), indicating that only a small portion of the original content is modified. And the embedding-level cosine similarity exceeds 0.9, demonstrating that the rewritten texts undergo minimal surface changes while maintaining high semantic consistency (see details and examples in appendix E).

Such minimal intervention supports the rationality of our mitigation strategies, which weaken shortcut patterns with-

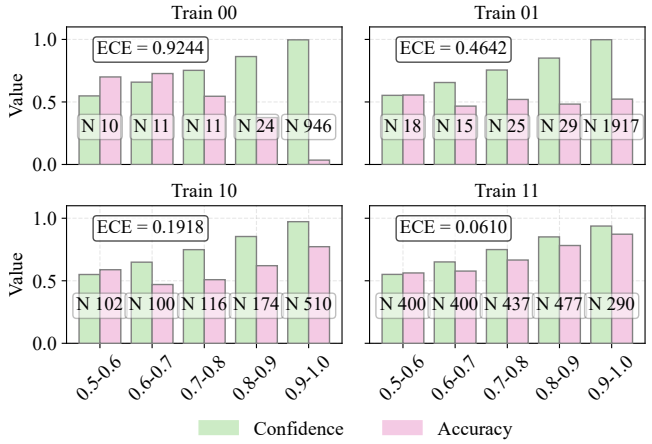

Figure 5: Calibration of detectors in standard and shortcut settings. The sub-caption "Train xy" indicates that the model is trained under setting x and tested under setting y, where 0 corresponds to the shortcut setting and 1 to the standard setting. The results demonstrate that models trained in the shortcut setting tend to exhibit higher confidence in their predictions, yet suffer from reduced accuracy.

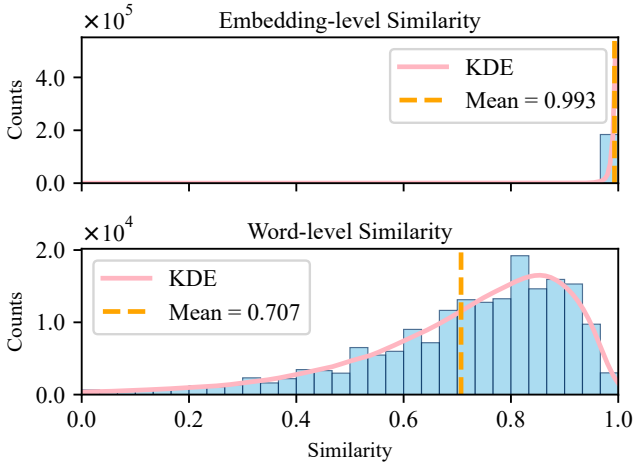

Figure 6: Distributions of embedding-level and token-level similarity scores. Cosine similarities at embedding-level mostly exceed 0.9, and edit similarities at token-level (computed as $1 - \frac{\text{edit distance}}{\text{raw tweet words}}$) are above 0.7 for the majority of text pairs, confirming that the rewriting process preserves semantic content while introducing minimal textual edits.

out distorting the core meaning and the underlying causal features relevant to classification.

5 Related Work

Shortcut Learning. Shortcut learning refers to the phenomenon where neural models exploit spurious correlations rather than truly understanding the underlying task (Geirhos et al. 2020; Du et al. 2023; Wan et al. 2025). This is a critical issue for ensuring the robustness of neural networks in real-world scenarios. Previous studies have made significant contributions in robustness of social bot detection. For example, Yang et al. (2020) investigated how to utilize prototypical samples to improve the robustness of detection models. Cresci et al. (2023) pointed out that some works exploit platform-specific spurious cues, such as verification status, to achieve better detection performance. Hays et al. (2023) highlighted the challenges of generalization across datasets in social bot detection. Recent research has also explored the use of dynamic information in the context of social bot detection (Zhou et al. 2023; He et al. 2024), or adopted more robust techniques such as unsupervised algorithms (Peng et al. 2024) to enhance generalization performance. Moreover, many recent studies have examined the increasing challenges of social bot detection in the era of LLMs (Ferrara 2023; Feng et al. 2024), and some have examined how content generated by LLMs can trigger shortcut learning risks(Wan et al. 2025). Collectively, these studies highlight the critical role of shortcut learning in the performance and generalization of social bot detection models.

Counterfactual Data Augmentation. CDA seeks to improve model robustness by generating examples whose labels flip under a target classifier. Early efforts in counterfactual data augmentation relied on human-authored rewrites or contrast sets (Kaushik, Hovy, and Lipton 2020; Gardner et al. 2020). Then various automated approaches have emerged, including rule-based perturbations (Ribeiro et al. 2020; Webster et al. 2020), control-guided generation (Madaan et al. 2021; Wu et al. 2021; Ross et al. 2022), and retrieval-based methods leveraging external knowledge (Paranjape, Lamm, and Tenney 2022). Recent work explores explanation-guided and distillation-based methods (Kim et al. 2023; Jeanneret, Simon, and Jurie 2024). With the rise of LLMs, prompt-based counterfactual generation has become increasingly popular (Madaan et al. 2023; Chiang and Lee 2023; Zheng et al. 2023; Kocmi and Federmann 2023; Mishra et al. 2024). We use LLMs to generate feature-specific counterfactual augmentation for capturing causal feature in social bot detection.

6 Conclusion

Our work reveals that shortcut features significantly affect social bot detection performance across multiple textual feature dimensions through triggering shortcut learning in training process. Building on counterfactual data augmentation using LLMs, we explore their effectiveness in addressing this challenging distribution shift from both the data and feature extraction perspectives. This work highlights the robustness gains achieved through this approach and offer new insights into designing more resilient social bot detectors. Moreover, our study provides a novel perspective on leveraging LLMs for causal interventions in scenarios with unknown or implicit ground truth. To the best of our knowledge, this is the first in-depth investigation into the impact of shortcut learning and the capacity for causal information extraction on social bot detection.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 62192781, No. 62272374), the Natural Science Foundation of Shaanxi Province (No. 2024JC-JCQN-62), the State Key Laboratory of Communication Content Cognition under Grant No. A202502, the Key Research and Development Project in Shaanxi Province (No. 2023GXLH-024), the Project of China Knowledge Center for Engineering Science and Technology, Huawei-Xi' an Jiaotong University Elite Class Program (No. INC-CHN2508012229) and the K. C. Wong Education Foundation.

References

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 178–186.

Chen, Y.; Zhang, Y.; Bian, Y.; Yang, H.; Kaili, M.; Xie, B.; Liu, T.; Han, B.; and Cheng, J. 2022. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35: 22131–22148.

Chiang, C.-H.; and Lee, H.-Y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631.

Cresci, S. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10): 72–83.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80: 56–71.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, 963–972.

Cresci, S.; Yang, K.-C.; Spognardi, A.; Di Pietro, R.; Menczer, F.; and Petrocchi, M. 2023. Demystifying misconceptions in social bots research. *Social Science Computer Review*, 08944393251376707.

Du, M.; He, F.; Zou, N.; Tao, D.; and Hu, X. 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1): 110–120.

Elmas, T.; Overdorf, R.; and Aberer, K. 2022. Characterizing retweet bots: The case of black market accounts. In *Proceedings of the international AAAI conference on web and social media*, volume 16, 171–182.

Feng, S.; Tan, Z.; Li, R.; and Luo, M. 2022. Heterogeneity-Aware Twitter Bot Detection with Relational Graph Transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4): 3977–3985.

Feng, S.; Wan, H.; Wang, N.; Li, J.; and Luo, M. 2021a. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 4485–4494.

Feng, S.; Wan, H.; Wang, N.; and Luo, M. 2021b. BotRGCN: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, 236–239.

Feng, S.; Wan, H.; Wang, N.; Tan, Z.; Luo, M.; and Tsvetkov, Y. 2024. What Does the Bot Say? Opportunities and Risks of Large Language Models in Social Media Bot Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3580–3601.

Ferrara, E. 2023. Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM*, 59(7): 96–104.

Gardner, M.; Artzi, Y.; Basmov, V.; Berant, J.; Bogin, B.; Chen, S.; Dasigi, P.; Dua, D.; Elazar, Y.; Gottumukkala, A.; et al. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1307–1323.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hays, C.; Schutzman, Z.; Raghavan, M.; Walk, E.; and Zimmer, P. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. In *Proceedings of the ACM web conference 2023*, 3660–3669.

He, B.; Yang, Y.; Wu, Q.; Liu, H.; Yang, R.; Peng, H.; Wang, X.; Liao, Y.; and Zhou, P. 2024. Dynamicity-aware Social Bot Detection with Dynamic Graph Transformers. In *IJCAI*.

Jeanneret, G.; Simon, L.; and Jurie, F. 2024. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4757–4767.

Kaushik, D.; Hovy, E.; and Lipton, Z. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Kim, S.; Oh, J.; Lee, S.; Yu, S.; Do, J.; and Taghavi, T. 2023. Grounding counterfactual explanation of image classifiers to textual concept space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10942–10950.

Kocmi, T.; and Federmann, C. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 193–203.

Li, W.; Deng, J.; You, J.; He, Y.; Zhuang, Y.; and Ren, F. 2025. ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation. In *Proceedings of the ACM on Web Conference 2025*, 4160–4170.

Liu, Q.; Kusner, M.; and Blunsom, P. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 187–197.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y.; Tan, Z.; Wang, H.; Feng, S.; Zheng, Q.; and Luo, M. 2023. Botmoe: Twitter bot detection with community-aware mixtures of modal-specific experts. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 485–495.

Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; and Zhang, G. 2019. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12): 2346–2363.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594.

Madaan, N.; Padhi, I.; Panwar, N.; and Saha, D. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 13516–13524.

Mannocci, L.; Mazza, M.; Monreale, A.; Tesconi, M.; and Cresci, S. 2024. Detection and characterization of coordinated online behavior: A survey. *arXiv preprint arXiv:2408.01257*.

Mishra, A.; Nayak, G.; Bhattacharya, S.; Kumar, T.; Shah, A.; and Foltin, M. 2024. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM Web Conference 2024*, 1538–1545.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paranjape, B.; Lamm, M.; and Tenney, I. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1670–1686.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3967–3976.

Peng, H.; Zhang, J.; Huang, X.; Hao, Z.; Li, A.; Yu, Z.; and Yu, P. S. 2024. Unsupervised social bot detection via structural information theory. *ACM Transactions on Information Systems*, 42(6): 1–42.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912.

Ross, A.; Wu, T.; Peng, H.; Peters, M. E.; and Gardner, M. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3194–3213.

Tardelli, S.; Nizzoli, L.; Tesconi, M.; Conti, M.; Nakov, P.; Da San Martino, G.; and Cresci, S. 2024. Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence. *Proceedings of the National Academy of Sciences*, 121(20): e2307038121.

Wan, H.; Wu, J.; Luo, M.; Zeng, Z.; and Su, Z. 2025. Truth over Tricks: Measuring and Mitigating Shortcut Learning in Misinformation Detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Wu, T.; Ribeiro, M. T.; Heer, J.; and Weld, D. S. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6707–6723.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 1096–1103.

Yang, Y.; Wu, Q.; He, B.; Peng, H.; Yang, R.; Hao, Z.; and Liao, Y. 2024. Sebot: Structural entropy guided multi-view contrastive learning for social bot detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3841–3852.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, M.; Zhang, D.; Wang, Y.; Geng, Y.-A.; and Tang, J. 2023. Detecting social bot on the fly using contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4995–5001.